Faculty of Information and Communication Technology

SMALL DATASET LEARNING IN PREDICTION MODEL USING BOX-WHISKER DATA TRANSFORMATION

Masitah binti Abdul Lateh

Master of Science in Information and Communication Technology

2020

SMALL DATASET LEARNING IN PREDICTION MODEL USING BOX-WHISKER DATA TRANSFORMATION

MASITAH BINTI ABDUL LATEH

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2020

DECLARATION

I declare that this thesis entitled “Small Dataset Learning in Prediction Model using Box-Whisker Data Transformation” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : …….

Name : Masitah Binti Abdul Lateh

Date : ……..

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature	:	……
Supervisor Name	:	Associate Professor Dr. Azah Kamilah Binti Draman @ Muda
Date	:	……

DEDICATION

I dedicate my master project to my beloved parents, Abdul Lateh Bin Man and Rohana Binti Zakaria, to my husband, Muhammad Kamal Ariffin Bin Mohd Aznan and to my siblings. I am really appreciated and thankful for your support towards completion of this project.

To both of my Supervisors, Associate Professor Dr. Azah Kamilah Binti Draman @ Muda and Associate Professor Dr. Mohd Sanusi Bin Azmi for the guidance and advice during the completion of my master project.

To all my beloved friends, for sharing and helping along the way.

ABSTRACT

There are several data mining tasks such as classification, clustering, prediction, summarization and others. Among them, a prediction task is widely applied in many real applications such as in manufacturing, medical, business and mainly for developing prediction model. However, to build a robust prediction model, the learning process from the training set are advised to have many samples. Otherwise, learning from small sample sizes might cause prediction task produced an imprecise model. However, to enlarge a sample size and ensure sufficient learning is sometimes difficult or expensive in certain situations. Thus, the information gained from small samples size are deficient. The main reason why a small sample size has problem in extracting the valuable information is that, the information gaps is exist. These gaps should be filled with observations in a complete dataset. However, these observations are not available. This situation has caused most of the learning tools are difficult to perform the prediction task. This is due to a small samples size will not provide sufficient information in the learning process which will lead to incorrect result. From the previous studies, there are solutions to improve learning accuracy and predictive capability where some artificial data will be added to the system using artificial data generation approach. Hence, the aims of this study are proposing an algorithm of hybrid to generate artificial samples adopts Small Johnson Data Transformation and Box-Whisker Plot which is introduced in previous studies. The proposed algorithm named as Box-Whisker Data Transformation considered all samples contain in a MLCC dataset in order to generate artificial samples. This study also investigates the effectiveness of employing the artificial data generation approach into a prediction model. Initially, the quantiles of raw samples are determine using Box-whisker Plot technique. Subsequently, the Small Johnson Data Transformation is employed to transformed raw samples to a Normal Distribution. Next, samples are generated from Normal Distribution. To test the effectiveness of the proposed algorithm, the real and generated samples is added to training phase to build a prediction model using M5 Model Tree. The results of this study are sample quantiles from reasonable range are generated. Not only that, using all samples available in a dataset as a training samples caused the properties of original pattern behaviors is retained. Besides, the effectivess of the learning performance of prediction model are proved when the number of artificial samples are increased, the average of the mean absolute Percentage Error (AvgMAPE) results of a M5 Model Tree are decreased. This reveals that the training size effect the accuracy of prediction models when the sample size is small.

ABSTRAK

Terdapat beberapa tugas 'data mining' seperti klasifikasi, klustering, ramalan, ringkasan dan lain-lain. Antara tugas-tugas ini, tugas ramalan sangat banyak digunakan dalam aplikasi sebenar seperti dalam pembuatan, perubatan, perniagaan dan terutamanya untuk menghasilkan model ramalan. Walau bagaimanapun, untuk membina sebuah model ramalan yang teguh, proses pembelajaran dari set latihan dinasihatkan supaya mempunyai sampel yang banyak. Jika tidak, pembelajaran daripada sampel yang kecil mungkin menyebabkan tugas ramalan menghasilkan model yang kurang tepat. Walau bagaimanapun, untuk membesarkan saiz sampel dan memastikan pembelajaran yang mencukupi, kadang-kala adalah sukar ataupun mahal dalam situasi tertentu. Oleh itu, maklumat yang diperoleh daripada sampel yang kecil adalah sedikit. Sebab utama mengapa saiz sampel yang kecil mempunyai masalah dalam mengekstrak maklumat yang berharga adalah disebabkan jurang maklumat telah wujud. Jurang ini harus diisi dengan 'observation' dalam dataset yang lengkap. Namun 'observation' ini tiada. Keadaan ini menyebabkan kebanyakan alat pembelajaran sukar untuk meramal. Ini adalah disebabkan saiz sampel yang kecil tidak akan memberikan maklumat yang mencukupi dalam proses pembelajaran yang akan membawa kepada keputusan ramalan yang salah. Daripada kajian yang terdahulu, terdapat penyelesaian untuk meningkatkan ketepatan pembelajaran dan keupayaan ramalan yang mana data buatan akan ditambah ke dalam sistem menggunakan pendekatan 'Artificial Data Generation'. Oleh itu, tujuan kajian ini adalah mencadangkan algoritma hibrid untuk menghasilkan sampel buatan menggunakan 'Small Johnson Data Transformation' dan 'Box-Whisker Plot' yang diperkenalkan dalam kajian terdahulu. Algoritma yang dicadangkan dinamakan sebagai 'Box-Whisker Data Transformation' menggunakan semua sampel dalam 'MLCC dataset' untuk menghasilkan sampel buatan. Kajian ini juga mengkaji keberkesanan pendekatann'Artificial Data Generation' ke dalam model ramalan. Pada mulanya, kuantil sampel ditentukan menggunakan teknik 'Box-Whisker Plot'. Seterusnya, 'Small Johnson Data Transformation' digunakan untuk mengubah sampel kepada Pengedaran Normal. Kemudian, sampel dihasilkan daripada Pengedaran Normal. Untuk menguji keberkesanan algoritma yang dicadangkan, sampel sebenar dan buatan dimasukkan ke dalam fasa latihan untuk membina model ramalan menggunakan M5 model tree. Hasil daripada kajian ini adalah kuantil sampel daripada julat yang munasabah dihasilkan. Bukan itu sahaja, menggunakan semua sampel yang ada dalam dataset sebagai contoh latihan menyebabkan sifat data yang asal dapat dikekalkan. Selain itu, kesan prestasi pembelajaran model ramalan dapat dibuktikan apabila bilangan sampel buatan meningkat, keputusan purata Ralat Peratusan Maksimum Mutlak (Avg MAPE) berkurangan. Ini mendedahkan bahawa saiz latihan memberi kesan kepada ketetapan model ramalan apabila saiz sampel kecil.

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my sincere acknowledgement to my Supervisors Associate Professor Dr. Azah Kamilah Binti Draman @ Muda and Associate Professor Dr. Mohd Sanusi Bin Azmi from the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM) for their essential supervision, support and encouragement towards the completion of this thesis. My deepest thanks for all help, supports, interests and valuable hints. Special thanks to UTeM's scholarship for the financial support throughout this project.

My deep appreciation is extended to my husband and my family, whose patient love also gives moral support which enabled me to complete this work. Without their blessings, I would never have this chance. Special thanks to all my friends and everyone who had been associated to the crucial parts of realization of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ANFIS	-	Adaptive Neuro-Fuzzy Inference System
ALH	-	Adaptive Hyperplane Algorithm
BWDT	-	Box Whisker Data Transformation
BPNN	-	Back-propagation Neural Network
CLTM	-	Central Location Tracking Method
DT	-	Decision Tree
GA	-	Genetic Algorithm
GEP	-	Genetic Programming
IQR	-	Interquartile Range
MAPE	-	Mean Absolute Percentage Error
MLCC	-	Multi-Layer Ceramic Capacitor
MLP	-	Multilayer Perceptron
MTD	-	Mega Trend Diffusion
NNWD	-	Neural Network Weight Determination
PID	-	Principle of Information Diffusion
PSO	-	Particle Swarm Optimization
QR	-	Quartile Range
SJDT	-	Small Johnson Data Transformation
SVM	-	Support Vector Machine

TSA	-	Trend Similarity Attribute
TTD	-	Tree Trend Diffusion
VSG	-	Virtual Sample Generation
WEKA	-	Waikato Environment for Knowledge Analysis

LIST OF PUBLICATIONS

1. Masitah, A. L., Azah, K. M., Zeratul, I. M. Y., Noor, A. M., and Sanusi, M. A., 2017. Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review. *Journal of Physics Conference Series,* 892(1), pp. 1-10.

CHAPTER 1

INTRODUCTION

1.1 Overview

Pattern recognition continues to be one of the key areas of research in the field of computer science and engineering. The sources of pattern recognition come from human, computer and data which is a data mining (Soman et al., 2005). The pattern recognition from data is a process of learning or observing the past data by studying the dependencies and extracting knowledge from data. The emergence of lots of new application has encouraged the pattern analysis and synthesis become significant subfield in pattern recognition. Generally, the pattern analysis analyzes patterns in data from the same source. The patterns obtained from the same source are then used to make the prediction of new data. In terms of Statistical Learning Theory, a pattern analysis produces a pattern based on a finite sample and provide bounds on the probability that pattern approximately represents a true pattern with some probability. However, the learning process usually required a large amount of training sample. Hence, the meaningful pattern of data is discovered and thus make the predictions purposes easier. Although in this recent years, a big data learning are getting attention, in certain situation like competitive manufacturing industries (Li et al., 2016b) and medical studies to diagnose related cancers (Chao et al., 2011), they faced difficulty in having large number of data samples. Thus, the uses of small samples in learning task is insufficient because the information extracted from samples might produce imprecise knowledge. This has been eventually getting attention from researchers to further explore to find solutions for effectively robust learning and accurate information from small dataset.

1.2 Research background

The amount of data in this world has been exploding and analyzing the big data will become a key basis of competition. The rapid growth of big data has leads to various applications in different industries such as in banking and securities, communication, media and services. While issues related to big data learning have only attracted attention in recent years, issues related to small-data learning have been revealed by the student's *t*-distribution in 1908 (Zabell, 2008; Li et al., 2018). Gaining the additional samples to enlarge a sample size and ensure the learning algorithm perform a sufficient learning is sometimes difficult or expensive (Hwe et al., 2018) in certain situations in different areas such as the diagnoses of rare disease (Huang et al., 2010; Chao et al., 2011), pattern recognition with limited pixel (Niyogi et al., 1998), development of new products (Li et al., 2013c; Li et al., 2016b) and automatic pedestrian parsing in surveillance video with limited pixel-label (Zheng et al., 2018).

Although machine learning algorithm such as decision trees (DT), genetic algorithm (GA), fuzzy theories and statistical method are extensively used in handling a small dataset problem, the extracted knowledge from small dataset is unreliable. This is because, majority of these algorithm were developed based on the assumptions that training sets can represent properties of populations. So, a knowledge extracted from small dataset to perform data mining tasks such as classification, clustering and prediction tasks may output less precise results for future events (Li et al., 2018). Among data mining tasks, prediction is one of the important tasks in data mining used in many areas. The prediction task involves developing a model based on the available data and this model is used in predicting future values of a new dataset of interest such as in weather forecasting. In medical area, the prediction model is used to predict which treatment would give the best solution for the problem using the molecular profiles and clinical information (Xiao et al.,

2014). Besides, in business, the companies that want to achieve their target market also can apply a prediction model to enable them to predict customer's demands on the future as early as possible.

The learning process of training samples is necessary to construct a prediction model. In the context of computational learning theory, the learning performances are affected when the sample size is small (Ruparel et al., 2013). This is because a small sample size fails to provide sufficient information because the gaps between two observations in small dataset exists. This situation is known as information gaps (Tsai and Li, 2015). These gaps should be filled with observations in a complete dataset. However, these observations are not available. Hence, the existed information gaps have caused most of the learning tools are difficult to predict (Tsai and Li, 2015) and fails to train their patterns with unavailable observations in the information gaps in small dataset (Li et al., 2018). Therefore, the availability of small sample size for analysis indirectly causes problems to real application to perform prediction task. This is because it failed to produce an accurate prediction from small sample size in learning process (Li et al., 2013d) and thus produce imprecise model.

Hence, when dealing with prediction model, it is advisable to have enough training samples to produce a reliable model. According to (Lin and Tsai, 2013) a sufficient training sample is defined as a sample set which provides enough information to a learning method to obtain stable learning accuracy.

Due to the large sample data is difficult to obtain in some areas, it is important to improve learning accuracy for small datasets obtained. From the previous studies, the most effective way to improve the accuracy of learning from the small datasets acquired as well as improving the predictive ability of machine learning algorithms is to add some artificial data to the system (Li et al., 2012b). The use of artificial data can be shown such as in

(Tsai and Li, 2008; Li et al., 2012b). They enlarge the amount of training data to be used in the training phase.

Literature shows that existing work in a small dataset problem is categorized into three main approaches namely data generation approaches, dimension reduction approaches and small dataset learning approaches. Artificial data generation approach applied fuzzy theories, decision tree, statistical method and other method. A dimension reduction used a feature selection and feature extraction method. Most of the existing techniques adopt artificial data generation approach to enlarge the amount of training set for the learning process before building a prediction model. By enlarge the amount of training set, the learning process is better and incomplete information in small dataset is filled.

Therefore, this study adopts the artificial data generation approach to enlarge training set by generating artificial samples. To demonstrate the effectiveness of the artificial samples, prediction task is used in this study where real and artificial samples are included in prediction techniques to build a prediction model and to improve the learning accuracy of prediction model. This study uses a real case study of multi-layered ceramics capacitors (MLCC) (Tsai and Li, 2008) which consist only small sample size that is difficult to produce an accurate prediction model.

The proposed hybrid algorithm of this study is a combination of two existing techniques proposed in previous studies to solve a small dataset problem. To enlarge the amount of training samples, the existing technique do not consider the entire samples in dataset but randomly choose samples as the first step to generate artificial samples. On the other hand, this study considered the entire samples to generate artificial samples.

1.3 Problem statement

A prediction model usually requires a large amount of sample size to perform a prediction task. However, in certain situations, a small sample size problem still arises which are considered as one of the important issues in prediction model (Li et al., 2013c). Hence, it is impractical to have a limited data size in building prediction model because a small sample size shows only limited properties and thus the knowledge extract is minimal. Thus, the suitable approach is required to increase the learning accuracy as well as learning stability of predictive models. Adding data to a system is an effective way to increase the predictive capability of machine learning algorithm through artificial data generation approach. This approach is widely used in handling a small dataset problem to enlarges the amount of training samples for learning purposes. By applying this approach, the information quality of training samples are improved and thus increasing the robustness of the learned model (Li et al., 2016a). In few studies, initially a sample was randomly selected (such as 6 and 10). However, a randomly choose samples approach to generate artificial samples cannot present the complete information of the sample distribution. Therefore, this study aimed to generates the artificial samples using the entire samples (44 samples) to retain the properties of original pattern behaviors. Although adding artificial data can be an effective approach when learning with small dataset, a former analysis was needed to infer the appropriate data distribution in which the data was generated (Li et al., 2016b). Therefore, this study was hybridizing these two methods that are Box-Whisker Plot and Small Johnson Data Transformation to successfully generate samples to be used as training set for learning purposes to create better prediction model. From literature, the Johnson's transformation is found as suitable method that able to bring data closer to normal distribution with satisfaction over a given statistical assumption. Usually, a sample size required for data transformation methods is usually large, and this motivates (Li et al.,

2016b) to develop new methods suitable for small dataset named as Small Johnson Data Transformation (SJDT). The advantage of applying this technique in this study is it able to convert small raw data to normal distribution to generate artificial samples.

1.4 Research questions

Based on the problem statement above, there are several research questions can be extracted.

1. Why hybrid techniques are required in generating artificial samples?
2. How to decide the suitable number of training sample to be used in learning process?
3. How to apply the proposed algorithm of Box-Whisker Data Transformation in prediction model to proof the effectiveness of the proposed algorithm?
4. How effective the hybridization technique in generating artificial sample?
5. How accurate is the proposed algorithm in prediction model?

1.5 Research objectives

The goal of this research is:

1. To develop an algorithm of Box Whisker Data Transformation for generating samples by hybridizing the Small Johnson Data Transformation and Box Whisker Plot techniques for small dataset.
2. To propose an approach of using all data samples in generating artificial samples for small dataset cases.
3. To validate the proposed algorithm by using a prediction model.